NeuroImage: Clinical 4 (2014) 718–729

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Longitudinal deformation models, spatial regularizations and learning strategies to quantify Alzheimer's disease progression

Jean-Baptiste Fiot [a,b,*,1], Hugo Raguet [b], Laurent Risser [d], Laurent D. Cohen [b], Jurgen Fripp [c], François-Xavier Vialard [b], for the Alzheimer's Disease Neuroimaging Initiative [2]

[a] IBM Research, Smarter Cities Technology Centre, Damastown, Dublin 15, Ireland
[b] CEREMADE, UMR 7534 CNRS, Université Paris Dauphine, PSL★, France
[c] CSIRO Preventative Health National Research Flagship ICTC, The Australian e-Health Research Centre — BioMedIA, Royal Brisbane and Women's Hospital, Herston, QLD, Australia
[d] CNRS, Institut de Mathématiques de Toulouse, UMR 5219, France

ARTICLE INFO

Article history:
Received 19 September 2013
Received in revised form 22 January 2014
Accepted 14 February 2014
Available online 1 April 2014

Keywords:
Alzheimer's disease
Brain imaging
Deformation model
LDDMM
Disease progression
Karcher mean
Transport
Logistic regression
Spatial regularization
Coefficient map

ABSTRACT

In the context of Alzheimer's disease, two challenging issues are (1) the characterization of local hippocampal shape changes specific to disease progression and (2) the identification of mild-cognitive impairment patients likely to convert. In the literature, (1) is usually solved first to detect areas potentially related to the disease. These areas are then considered as an input to solve (2). As an alternative to this sequential strategy, we investigate the use of a classification model using logistic regression to address both issues (1) and (2) simultaneously. The classification of the patients therefore does not require any a priori definition of the most representative hippocampal areas potentially related to the disease, as they are automatically detected. We first quantify deformations of patients' hippocampi between two time points using the *large deformations by diffeomorphisms* framework and transport these deformations to a common template. Since the deformations are expected to be spatially structured, we perform classification combining logistic loss and *spatial regularization* techniques, which have not been explored so far in this context, as far as we know. The main contribution of this paper is the comparison of regularization techniques enforcing the coefficient maps to be spatially smooth (Sobolev), piecewise constant (total variation) or sparse (fused LASSO) with standard regularization techniques which do not take into account the spatial structure (LASSO, ridge and ElasticNet). On a dataset of 103 patients out of ADNI, the techniques using spatial regularizations lead to the best classification rates. They also find coherent areas related to the disease progression.

1. Introduction

Large scale population studies aim to improve the understanding of the causes of diseases, define biomarkers for early diagnosis, and develop preventive treatments. An important challenge for medical imaging is to analyze the variability in MRI acquisitions of normal control (NC), mild cognitive impairment (MCI), and Alzheimer's disease (AD) patients. For Alzheimer's disease, several classification strategies have been proposed to separate patients according to their diagnosis. These methods can be split into three categories: voxel-based (Fan et al., 2007, 2008a,b; Klöppel et al., 2008; Lao et al., 2004; Magnin et al., 2009; Vemuri et al., 2008), cortical-thickness-based (Desikan et al., 2009; Klöppel et al., 2008; Querbes et al., 2009) and hippocampus-based (Chupin et al., 2007, 2009; Gerardin et al., 2009) methods. While decent classification rates can be achieved to separate AD from NC or NC from p-MCI (progressive MCI patients, i.e. converting to AD), all methods perform poorly at separating s-MCI (stable MCI patients, i.e. non-converting to AD) and p-MCI. A recent review comparing these methods can be found in Cuingnet et al. (2011).

In the case of longitudinal analysis, it is not anymore the shapes that are compared but their evolutions in time. To extract information between two successive time-points, we use a one-to-one deformation which maps the first image onto the second one. Different registration algorithms are available to compute plausible deformations in this context. However, only one, the *large deformations via diffeomorphisms (LDDMM)* (Beg et al., 2005), provides a Riemannian setting that enables to represent the deformations using tangent vectors: initial velocity

* Corresponding author at: IBM Research, Smarter Cities Technology Centre, Damastown, Dublin 15, Ireland.
E-mail address: jean-baptiste.fiot@centraliens.net (J.-B. Fiot).

[1] Now affiliated to IBM Research. The work and preparation of this article were mostly done while being affiliated to Université Paris Dauphine.

[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

http://dx.doi.org/10.1016/j.nicl.2014.02.002

fields or equivalently initial momenta. This can be used in practice to retrieve local information and to perform statistics on it as presented in Vaillant et al. (2004) and Wang et al. (2007). In this direction, it is worth mentioning the study of Singh et al. (2010) which shows the correlation between principal modes of deformation and diagnosis. In Qiu et al. (2008) the authors estimate the typical deformation of several clinical groups from the deformations between baseline and follow-up hippocampus surfaces. In order to compare this information across the population, we need to define a common coordinate system. This implies (1) the definition of a template and (2) a methodology for the transport of the tangent vector information. Note finally that, as far as the authors know, no paper explores binary classification using logistic regression in this context.

Quality of shape descriptors with regard to the disease is often evaluated through statistical significance tests or classification performance. In this paper, we evaluate descriptors on a binary classification task using logistic regression.

In addition to its simplicity, it has the advantage of providing a map of coefficients weighting the relevance of each voxel. Such map can be used to localize the hippocampus deformations that are related to AD. However, the dimensionality of the problem (i.e. number of voxels p) being much higher than the number of observations (i.e. number of patients n, $p \sim 10^6 \gg n \sim 10^2$), the problem requires proper regularization.

Now standard regularization methods such as *ridge* (Hoerl and Kennard, 1970), *LASSO* (Tibshirani, 1994) and *Elastic Net* (Zou and Hastie, 2005) do not take into account any spatial structure of the coefficients.

In contrast, spatial models for regularizing supervised learning methods have been proposed in the literature (Grosenick et al., 2013; Jenatton et al., 2012; Ng and Abugharbieh, 2011). *Total variation* was used to regularize a logistic regression on functional MRI (fMRI) data (Michel et al., 2011). This method promotes coefficient maps with spatially homogeneous clusters. *Fused LASSO* was also used on fMRI data (Baldassarre et al., 2012; Gramfort et al., 2013). Similar ideas can be found in Cuingnet et al. (2012) where the authors defined the notion of spatial proximity to regularize a linear SVM classifier.

In Durrleman et al. (2013), the authors introduce sparse parametrization of the diffeomorphisms in the LDDMM framework. Our goal is different: we want spatial properties (smoothness, sparsity, etc.) to be found *across* the population (i.e. on the common template) and we want this coherence to be driven by the disease progression.

In this paper, we investigate the use of total variation, Sobolev and fused LASSO regularizations in 3D volumes. Compared to total variation, Sobolev enforces smoothness of the coefficient map, whereas fused LASSO adds a sparsity constraint.

The deformation model used to assess longitudinal evolutions in the population is presented in Section 2. Machine learning strategies are discussed and the model of classification with logistic loss and spatial regularization is described in Section 3. The dataset used and numerical results are presented in Section 4. We illustrate that initial momenta capture information related to AD progression, and that spatial regularizations significantly increase classification performance. Section 5 concludes the paper.

2. Longitudinal deformation model for population analysis

2.1. Global pipeline

Let us assume that we have a population of patients and the binary segmentation of their hippocampus at two different time points, called *screening* and *follow-up*. Let us also assume that all patients have the same diagnosis at the screening time point, and only a part of them have converted to another diagnosis at the follow-up time point. Our goal is to compare patient evolutions, and classify them with regard to disease progression, i.e. stable diagnosis versus progressive diagnosis. From a machine learning point of view, we need to build features encoding the evolutions of the patients.

We use the pipeline summarized in Fig. 1. First, the evolution descriptors are computed locally for each patient (independently). To be able to compare these descriptors, one needs to transport them into a common space. To do so, a population template is computed, towards which all the local descriptors are transported. Finally, classification is performed to separate progressive from stable patients.

2.2. Diffeomorphic registration via geodesic shooting

As mentioned in Sections 1 and 2.1, local deformation descriptors are computed to model the evolutions of the patients. In this section, we describe how we use diffeomorphic registration via geodesic shooting Vialard et al. (2012a) to compute these local deformation descriptors.

2.2.1. Definitions

To register a source image $I : \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ towards a target image $J : \Omega \subset \mathbb{R}^3 \to \mathbb{R}$, the LDDMM framework (Beg et al., 2005) introduces the following minimization problem

$$\underset{v \in L^2([0,1],\mathcal{H}_K)}{\operatorname{argmin}} \frac{1}{2}\|I \circ \phi_{0,1}^{-1} - J\|_{L^2}^2 + \lambda \int_0^1 \|v_t\|_K^2 dt, \tag{1}$$

where $v : (t,\omega) \in [0,1] \times \Omega \subset \mathbb{R}^3 \to \Omega$ is a time dependent velocity field that belongs to a reproducing kernel Hilbert space $\mathcal{H}_K$ of smooth enough vector fields defined on Ω, and of associated kernel K and norm $\|\ \|_K$, and $\lambda \geq 0$ is a regularization coefficient. For $(t,\omega) \in [0,1] \times \Omega$, we note $v_t(\omega) = v(t,\omega)$. The deformation $\phi : [0,1]^2 \times \Omega \subset \mathbb{R}^3 \to \Omega$ is given by the flow of v_t

$$\forall (t,\omega) \in [0,1] \times \Omega, \quad \begin{cases} \dfrac{\partial \phi_{0,t}}{\partial t}(\omega) = v_t \circ \phi_{0,t}(\omega) \\ \phi_{t,t}(\omega) = \omega\ , \end{cases} \tag{2}$$

where $\phi_{t1,\,t2}$ is the deformation from $t = t_1$ to $t = t_2$. Such approach induces a Riemannian metric on the orbit of I, i.e. the set of all deformed images by the registration algorithm (Miller et al., 2006). The first term in formula (1) is a similarity term controlling the matching quality whereas the second one is a smoothing term controlling the deformation regularity. Now noting $I_t \overset{\text{def.}}{=} I \circ \phi_{0,t}^{-1}$ and $J_t \overset{\text{def.}}{=} J \circ \phi_{t,1}$,

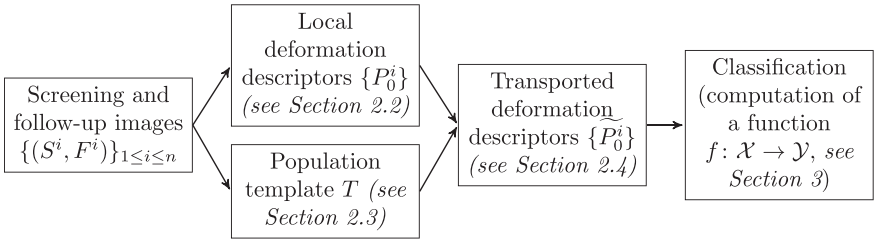


Fig. 1. Four steps are needed to classify patient evolutions using local descriptors of shape deformations: (1) the local descriptors are computed for each patient independently, (2) a population template is computed, (3) all local shape deformation descriptors are transported towards this template, and (4) classification is performed.

the Euler-Lagrange equation associated with Eq. (1) reads $\forall\ (t,\omega) \in [0,1] \times \Omega$,

$$v_t(\omega) = -K \star \Big(\text{grad}\, I_t(\omega)\, \text{Jac}_{\phi_{t,1}}(\omega)(I_t(\omega) - J_t(\omega))\Big), \tag{3}$$

where K is the translation-invariant kernel of the reproducing kernel Hilbert space, $\star$ the convolution operator, grad the image gradient in space and Jac_ϕ the Jacobian of ϕ.

For $t \in [0,1]$, let us define the momentum $P_t : \Omega \to \mathbb{R}$ by

$$\forall \omega \in \Omega, \quad P_t(\omega) \overset{\text{def.}}{=} \text{Jac}_{\phi_{t,1}}(\omega)(I_t(\omega) - J_t(\omega)). \tag{4}$$

The Euler–Lagrange Eq. (3) can be rewritten as a set of geodesic shooting equations

$$\forall (t,\omega) \in [0,1] \times \Omega, \quad \begin{cases} \dfrac{\partial I_t}{\partial t}(\omega) + \langle\, \text{grad}\, I(\omega), v_t(\omega)\, \rangle = 0, \\ \dfrac{\partial P_t}{\partial t}(\omega) + \text{div}\,(P_t(\omega) v_t(\omega)) = 0, \\ v_t(\omega) + K \star \text{grad}\, I_t(\omega)\ P_t(\omega) = 0, \end{cases} \tag{5}$$

where div is the divergence operator.

Given an initial image I_0 and an initial momentum P_0, one can integrate the system (Eq. (5)). Such a resolution is called *geodesic shooting*. We say that *we shoot from* I_0 *using* P_0.

The minimization problem (Eq. (1)) can be reformulated using a shooting formulation on the initial momentum P_0

$$\underset{P_0}{\text{argmin}}\ \frac{1}{2} \| I \circ \phi_{0,1}^{-1} - J \|_{L^2}^2 + \lambda \langle \text{grad}\, I_0 P_0, K \star \text{grad}\, I_0 P_0 \rangle_{L^2} \tag{6}$$

subject to the shooting system (Eq. (5)).

In order to solve the new optimization problem (Eq. (6)), we use the methodology described in Risser et al. (2011) and Vialard et al. (2012a). Note that this methodology is similar to the one presented in Ashburner and Friston (2011), but uses a different optimization strategy.

For each patient, a two-step process was performed to encode the deformations of the hippocampus shape evolution from the screening image S (scanned at $t = t_0$) to the follow-up image F (scanned at $t = t_0 + 12$ months), as described in Fig. 2. First F was rigidly registered back to S. We note $R : \Omega \subset \mathbb{R}^3 \to \Omega$ the rigid transformation obtained. Second, the geodesic shooting was performed with the screening image as source image ($I = S$) point towards the registered followed-up image as target image ($J = F \circ R^{-1}$). Initial momenta from different patients are local descriptors that were used to compare hippocampus evolutions, such choice is further described in the next paragraph.

2.2.2. Motivation and rationales for the use of initial momenta

As written in the third row of Eq. (5), the velocity field v encoding the geodesic between the registered images has the following property at each time $t \in [0,1]$ and at each coordinate $\omega \in \Omega$,

$$v_t(\omega) = -K \star \text{grad}\, I_t(\omega) P_t(\omega), \tag{7}$$

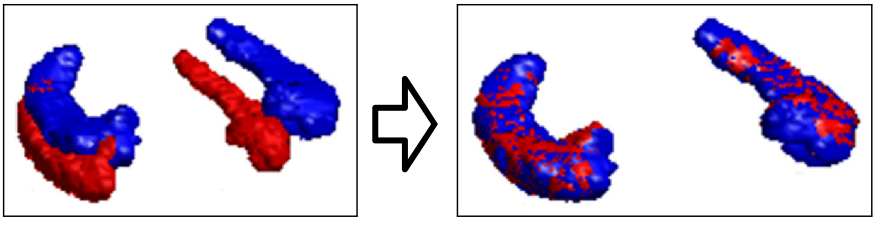
We recall that I_t, v_t and P_t are respectively the deformed source image, the velocity field and the momentum at time t. We also denote $K\star$ the convolution with the kernel K (typically Gaussian). Therefore, Eq. (7) can be read in the case of a binary image as follows: the unitary vector field normal to the shape surface is multiplied by a scalar field $P(t)$ and this quantity gives the vector field v_t once convolved with the kernel K.

The system given in all rows of Eq. (5) leads to the fact that the initial momentum P_0 entirely controls the deformation for a given source image I_0 and a given kernel K. In the context of our study, longitudinal variations of the geodesics are relatively limited as only small deformations are required to register pairs of hippocampi out of the same subject. The displacement field can then be reasonably approximated by $Id + v_0$ using a first-order expansion of Eq. (5). As a consequence, P_0 can be directly interpreted as a value encoding expansions and contractions of the shape if multiplied by $-\text{grad}\, I_0$ and then smoothed by K. Note also that the momentum is a scalar field, which is a more compact representation than a vector field. This motivates our approach.

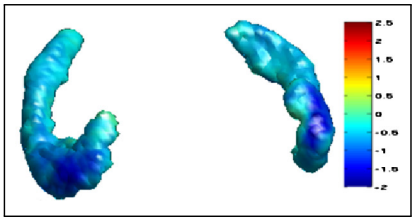
2.3. Population template

2.3.1. Need for a template

As mentioned in Section 2.1, local descriptors of hippocampus evolutions need to be transported in a common space prior to any statistical analysis. One way to obtain spatial correspondences between local descriptors of different patients consists in building a population template and then aligning these descriptors on the template. In the literature, template algorithms can be categorized into deterministic (Avants and Gee, 2004; Beg and Khan, 2006; Fletcher et al., 2004; Pennec, 2006; Vialard et al., 2011), probabilistic (Allassonnière et al., 2008; Ma et al., 2008) and mixed (Bhatia et al., 2004; Jia et al., 2010; Joshi et al., 2004; Seghers et al., 2004) approaches. As described in Section 4.1, we want



a) Step 1: rigid registration.



b) Step 2: geodesic shooting.

Fig. 2. For each patient, the initial momentum encoding the hippocampus evolution is computed in a two-step process.

to build a population of binary images of hippocampi. As there is no variation of topology, and we want a template with sharp boundaries without averaging the gray levels, the first category is appropriate. Most algorithms in this category rely on the notions of Fréchet and Karcher means, which we will now describe.

2.3.2. Notions of Fréchet and Karcher means

In the Riemannian framework used for the geodesic shooting, a Fréchet mean (Fréchet, 1948) can be used to define an average shape from a population (Fletcher et al., 2004; Pennec, 1999, 2006). Given n images $\{S^i : \Omega \subset \mathbb{R}^3 \to \mathbb{R}\}_{1 \le i \le n}$ and d a Riemannian metric on the space of images, the *Fréchet mean* $\hat{T} : \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ is defined as a minimizer of the sum of the geodesic distances to all images

$$\min_T \quad \frac{1}{n}\sum_{i=1}^{n} d\left(T, S^i\right)^2. \tag{8}$$

In practice, such problem is often solved via an optimization procedure looking for a local minimum, and the solutions found are called *Karcher means*. For instance, a solution of Eq. (8) can be computed using a gradient descent procedure (Vialard et al., 2011).

2.3.3. Invariance to rigid orientations, approximations and optimization procedure

The problem (8) is not invariant with respect to the rigid orientations of the input images, we modify the optimization problem to

$$\min_{T, R^1, \ldots, R^n} \quad \frac{1}{n}\sum_{i=1}^{n} d\left(T, S^i \circ \left(R^i\right)^{-1}\right)^2, \tag{9}$$

where $\{R^i : \Omega \to \Omega\}_{1 \le i \le n}$ are rigid transformations. In this paper, we assume that the solution of Eq. (9) can be approximated by alternate minimization. It is also important to note that in the general case there is not necessarily unicity of the solution.

When the $\{R^i\}$ are fixed, we follow the optimization strategy described in Vialard et al. (2011). Since the functional in Eq. (1) does not give a geodesic distance between two images — but between a source image and the deformed image, we approximate the minimization with regard to T by

$$\min_T \quad \frac{1}{n}\sum_{i=1}^{n} d\left(T, J_1^i\right)^2, \tag{10}$$

where J_1^i is the result of the shooting equations for the initial conditions $I = T$ and $P_0 = P_0^i$, where P_0^i is a minimizer of Eq. (6) with $J = S^i \circ (R^i)^1$. In this case, each term of the sum in Eq. (10) is equal to $<\mathrm{grad} I P_0, K \star (\mathrm{grad} I P_0)>_{L^2}$, and the gradient with regard to T is

$$-\frac{1}{n}\sum_{i=1}^{n} K \star \mathrm{grad}\, T P_0^i, \tag{11}$$

where P_0^i is the initial momentum matching T on $S^i \circ (R^i)^{-1}$ via the shooting system (Eq. (5)).

When T is fixed, we approximate the optimization over $\{R^i\}_{1 \le i \le n}$ by performing rigid registrations matching each S^i to T.

Altogether, each update of the Karcher estimate is composed of four steps

1. the images S^i are rigidly aligned towards the current Karcher mean estimate T_k,
2. diffeomorphic registrations via geodesic shootings from the current Karcher estimate T_k towards all the registered images $S^i \circ (R^i)^{-1}$ are computed,
3. geodesic shooting from T_k using $P_0^{mean} \overset{\text{def.}}{=} \frac{1}{n}\sum_i P_0^i$ generates a deformation field u_{mean},
4. the composed deformation field $u_{k+1} \overset{\text{def.}}{=} u_{mean} \circ u_k$ is used to compute the updated estimate from the reference image.

The advantage of computing the new estimate from a reference image is to avoid consecutive resamplings that would lead to smoothing and bias, as noted in Yushkevich et al. (2010).

In the literature, the empirical convergence of the gradient descent procedure optimizing over T (with $\{R_i\}_{1 \le i \le n}$ fixed) was studied in Vialard et al. (2011, 2012b). Similar tests are performed in Section 4.2 for our procedure.

2.4. Tangent information and associated transport

2.4.1. Motivation and rationals

The local descriptors computed for each patient as explained in Section 2.2 need to be transported in a common coordinate space: the space of the Karcher average defined in Section 2.3.

There is still no consensus about the choice of which transport method should be used in our context. Different methods have been proposed. The first one is the transport of vector fields by the standard adjoint map. It was however shown that this method is not quite appropriate for statistical study (Bossa et al., 2010). Parallel transport was also proposed in the context of LDDMM (Younes, 2007). Although it might seem relevant in our context, volume variation may be distorted. Note that its properties also depend on the deformation path and not only on the final deformation.

In the context of LDDMM, another action of the group of deformations on the momentum is called co-adjoint transport (Fiot et al., 2012). This method only depends on the final deformation and preserves volume variation in the context of small deformations on binary images. This argument motivated its use in our study.

2.4.2. Definitions

A two-step process was then used to transport local descriptors of hippocampus evolutions to the template space (Fig. 3). First, the screening hippocampus S^i was registered towards the template T rigidly (Ourselin et al., 2001) then non-rigidly (Modat et al., 2010). The resulting deformation is denoted by ϕ^i. Second, this transformation was used to transport the local descriptors of hippocampus deformations towards the template.

We use the standard transport for a density $P_0^i : \Omega \subset \mathbb{R}^3 \to \mathbb{R}$, defined by

$$\forall \omega \in \Omega, \quad \widetilde{P}_0^i(\omega) \overset{\text{def.}}{=} \det\left(\mathrm{Jac}_{\left(\phi^i\right)^{-1}(\omega)}\right) P_0^i \circ \left(\phi^i\right)^{-1}(\omega), \tag{12}$$

where det is the notation for the determinant. Note that this action preserves the global integration of the density by a simple change of variable.

3. Machine learning strategies

3.1. Support vector machine classification

In Fiot et al. (2012), SVM classifiers are used on different types of features. In that paper, local features obtained by integration of initial momenta on subregions provided the best classification results. This conclusion motivates the search for an optimal subregion Ω_r defining features as $\mathbf{x}_i \overset{\text{def.}}{=} \int_{\Omega_r} \widetilde{P}_0^i(\omega) d\omega$ (optimal in terms of classification accuracy). This is equivalent to the search of the best indicator function $I_r : \Omega \to \{0,1\}$, or more generally a weighting function $w : \Omega \to \mathbb{R}$ defining features by $\mathbf{x}_i \overset{\text{def.}}{=} \int_\Omega w(\omega) \widetilde{P}_0^i(\omega) d\omega$.

To compute meaningful weighting functions, models where the *feature space* is the same as the *input space* are of particular interest. Indeed as one coefficient corresponds to one voxel, meaningful spatial

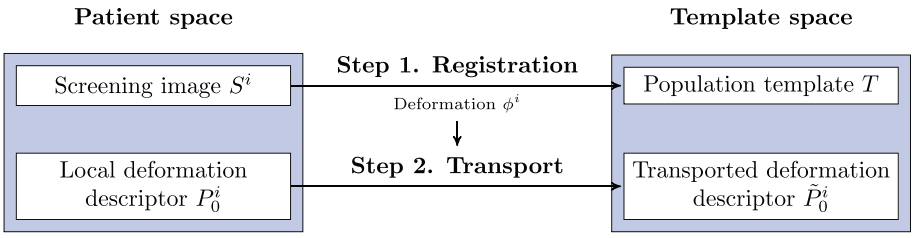


Fig. 3. Local descriptors of hippocampus evolutions are transported to the template in a two-step process. First the deformation field from the patient space to the population template. Second, this deformation field is used to transport the local descriptors.

regularizations can be introduced. This was used in the linear SVM setting in Cuingnet et al. (2012). In this paper, we exploit similar ideas on a classification framework with a logistic loss, which is well-suited for the introduction of spatial regularizations, easy to implement and that can be solved efficiently.

3.2. Binary classification with logistic regression and spatial regularization

3.2.1. Definitions

Let us define a predictive model which reads

$$\mathbf{y} \overset{\text{def.}}{=} F(\mathbf{X}\mathbf{w} + b), \tag{13}$$

where $\mathbf{y} \in \{\pm 1\}^n$ is the behavioral variable, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix containing n observations of dimension p, F is the prediction function and $(\mathbf{w},b) \in \mathbb{R}^p \times \mathbb{R}$ are the parameters to estimate. In our application, each coefficient in $\mathbf{y}$ represents the disease progression of one of the n patients, and each row in $\mathbf{X}$ contains the initial momentum representing the deformations of the hippocampus of one of the n patients. It is important to notice that each row in $\mathbf{X}$ is noted as a vector in $\mathbb{R}^p$ in the formulation of the predictive model, but it is actually a scalar field in 3D. Similarly, $\mathbf{w}$ is noted as a vector in $\mathbb{R}^p$ for the convenience of the formulation of the model, even if it also represents a scalar field in 3D. Since each coefficient in $\mathbf{w}$ is associated to a spatial position, $\mathbf{w}$ is sometimes called a *coefficient map*. Such property allows us to detect (spatial) areas of interest, with regard to the machine learning problem we want to solve (see Section 3.2.4 about the interpretation of the solution of the model).

The logistic regression model defines the probability of observing y_i given the data $\mathbf{x}_i$ as

$$p(y_i|\mathbf{x}_i, \mathbf{w}, b) \overset{\text{def.}}{=} \frac{1}{1 + \exp(-y_i(\mathbf{x}_i^T\mathbf{w} + b))}. \tag{14}$$

Given parameters $\left(\hat{\mathbf{w}}, \hat{b}\right)$ and a new data point $\mathbf{x}$ the prediction is the maximum likelihood, i.e. $\text{class}(\mathbf{x}) = \text{argmax}_{y \in \{\pm 1\}} p\left(y|\mathbf{x}, \hat{\mathbf{w}}, \hat{b}\right) = \text{sign}\left(\mathbf{x}^T\hat{\mathbf{w}} + \hat{b}\right)$. Accordingly the parameters are estimated as minimizers of the opposite log likelihood of the observations, considered as independent

$$\mathcal{L}(\mathbf{w}, b) \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i\left(\mathbf{x}_i^T\mathbf{w} + b\right)\right)\right). \tag{15}$$

Since the number of observations is much smaller than the dimension of the problem ($n \ll p$) minimizing directly the loss Eq. (15) leads to overfitting, and proper regularization is required. This is commonly performed by introducing a regularization function J and the final problem becomes

$$\text{Find}\left(\hat{w}, \hat{b}\right) \text{in} \underset{w,b}{\text{argmin}}\ \mathcal{L}(w, b) + \lambda J(w), \tag{16}$$

where λ is a coefficient tuning the balance between loss and regularization.

The standard *elastic net* regularization (Zou and Hastie, 2005) uses a combined ℓ_1 and squared ℓ_2 penalization $\lambda EN(\mathbf{w}) \overset{\text{def.}}{=} \lambda_1||\mathbf{w}||_1 + \lambda_2||\mathbf{w}||_2^2 = \sum_{j=1}^{p}\lambda_1|w_j| + \lambda_2 w_j^2$, with the limit cases $\lambda_2 = 0$ referred to as *LASSO* (Tibshirani, 1994) and $\lambda_1 = 0$ referred to as *ridge* (Hoerl and Kennard, 1970). However as mentioned in Michel et al. (2011), one drawback of such methods is that they do not take into account any geometrical structure of $\mathbf{w}$. Since coefficients are expected to be locally correlated in space, we investigate the Sobolev semi-norm, total variation semi-norm and fused-LASSO regularizations, respectively defined as

$$\text{SB}(w) \overset{\text{def.}}{=} \sum_{\omega \in \Omega} \|\text{grad}_\Omega w(\omega)\|_2^2, \tag{17}$$

$$\text{TV}(\mathbf{w}) \overset{\text{def.}}{=} \sum_{\omega \in \Omega} \|\text{grad}_\Omega \mathbf{w}(\omega)\|_2, \tag{18}$$

$$\lambda\text{FL}(\mathbf{w}) \overset{\text{def.}}{=} \lambda_1 \text{TV}(\mathbf{w}) + \lambda_2\|\mathbf{w}\|_1. \tag{19}$$

The above sums go over all voxels ω in the domain $\Omega \subset \mathbb{R}^3$, and grad_Ω is a linear operator implementing the image gradient by finite differences. By indexing each voxel ω by integer coordinates on a 3D lattice, we define grad_Ω by

$$\text{grad}_\Omega \mathbf{w}\left(\omega_{ijk}\right) \overset{\text{def.}}{=} \begin{pmatrix} \Delta_\Omega \mathbf{w}\left(\omega_{ijk}, \omega_{(i+1)jk}\right) \\ \Delta_\Omega \mathbf{w}\left(\omega_{ijk}, \omega_{i(j+1)k}\right) \\ \Delta_\Omega \mathbf{w}\left(\omega_{ijk}, \omega_{ij(k+1)}\right) \end{pmatrix}, \tag{20}$$

where $\Delta_\Omega \mathbf{w}(\omega_1, \omega_2) \overset{\text{def.}}{=} \begin{cases} \mathbf{w}(\omega_2) - \mathbf{w}(\omega_1) & \text{if } (\omega_1, \omega_2) \in \Omega^2, \\ 0 & \text{otherwise.} \end{cases}$ This definition allows to restrain Ω to any region of interest and boundaries of the domain are not penalized. Rationals and differences for those regularizations are discussed in Section 4.

3.2.2. Solving the model

Let us first study differentiability and convexity of the objective function in Eq. (15). For convenience, we define $\widetilde{\mathbf{w}} \overset{\text{def.}}{=} (\mathbf{w}^T, b)^T$ and for all i, $\widetilde{\mathbf{x}}_i \overset{\text{def.}}{=} (\mathbf{x}_i^T, 1)^T$, with associated data matrix $\widetilde{\mathbf{X}} \overset{\text{def.}}{=} (\widetilde{\mathbf{x}}_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le p+1}} \in \mathbb{R}^{n \times (p+1)}$. Then Eq. (15) becomes

$$\mathcal{L}(\widetilde{\mathbf{w}}) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i\widetilde{\mathbf{x}}_i^T\widetilde{\mathbf{w}}\right)\right). \tag{21}$$

This loss function is twice differentiable and the non-negativity of $\nabla^2\mathcal{L}(\widetilde{\mathbf{w}})$ establishes the convexity.

When the regularization J is also convex and twice differentiable the reference optimization algorithms include quasi-Newton methods; in particular for large-scale problems the limited memory Broyden–Fletcher–Goldfarb–Shanno (LM-BFGS) is very popular. However non-differentiable regularizations such as total variation and fused LASSO optimization raises theoretical difficulties. Proximal methods such as monotonous fast iterative shrinkage thresholding algorithm (M-FISTA,

(Beck and Teboulle, 2009)) and generalized forward–backward (GFB, (Raguet et al., 2013)) have been considered. Unfortunately their low convergence rates are prohibitive for extensive investigation of the classification scheme (parameter λ, domain Ω, training design matrix $\mathbf{X}$). Therefore we used the hybrid algorithm for non-smooth optimization (HANSO, (Lewis and Overton, 2012)) which is a LM-BFGS algorithm with weak Wolfe conditions line search. This addresses both the total variation semi-norm and the ℓ_1-norm, with almost everywhere

$$\nabla \mathrm{TV}(\mathbf{w}) = -\mathrm{div}\left(\left(\|\mathrm{grad}_\Omega \mathbf{w}(\omega)\|_2^{-1}\mathrm{grad}_\Omega \mathbf{w}(\omega)\right)_{\omega\in\Omega}\right),$$
$$\nabla\|\mathbf{w}\|_1 = (\mathrm{sign}(\mathbf{w}(\omega)))_{\omega\in\Omega}.$$

3.2.3. Weighted loss function

In supervised learning, classifiers trained with observations not equally distributed between classes can be biased in favor of the majority class. In order to alleviate this, several strategies can be used. One strategy is to restrict the training set to be equally distributed among classes. An alternative strategy is to use the full training set and introduce weights $(q_i)_{i\in[[1,n]]}$ in the loss function as follows

$$\mathcal{L}_q(\tilde{\mathbf{w}}) \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} q_i \log\left(1+\exp\left(-y_i\,\tilde{\mathbf{x}}^T\,\mathbf{w}\right)\right) \tag{22}$$

where $q_i \overset{\text{def.}}{=} n/\left(n_c \times \mathrm{card}\left\{j\in[1\,..n] | y_j = y_i\right\}\right)$, n_c being the number of classes (2 in our case). When the observations are equally distributed among classes $q_i = 1$ for all i and one retrieves (Eq. (21)), whereas $q_i < 1$ (respectively $q_i > 1$) when the class of observation i is over-represented (respectively under-represented) in the training set.

3.2.4. Interpretation of the solution

Another motivation for the use of the model presented in Section 3 is the possibility to interpret the computed solution. Let us remind that, after optimization, the solution is of the form $\left(\hat{\mathbf{w}},\hat{b}\right)\in\mathbb{R}^p\times\mathbb{R}$. This solution can be used to predict the evolution $y\in\{\pm 1\}$ of a new patient of associated initial momentum $\mathbf{x}\in\mathbb{R}^p$, by using the equation $y = \mathrm{sign}\left(\mathbf{x}^T\hat{\mathbf{w}}+\hat{b}\right)$. As mentioned in Section 3.2.1, the hyperplane $\hat{\mathbf{w}}$ has the same dimension of the initial momentum, and each coefficient is associated to one voxel.

Now let us talk about the interpretation of the weights in $\hat{\mathbf{w}}$. High coefficients in $\hat{\mathbf{w}}$ correspond to *areas of the hippocampus where the deformation is related to the disease progression*. They are not areas of high expansions or contractions, and therefore have a different interpretation than the coefficients in the initial momenta (see Section 2.2 for the interpretation of the coefficients of the initial momenta). On the contrary, coefficients close to zero in $\hat{\mathbf{w}}$ represent areas where the values of $\mathbf{x}$ are not relevant to the disease progression (in that case the values of $\mathbf{x}$ in these areas will not modify the value of the scalar product $\mathbf{x}^T\hat{\mathbf{w}}$). In that sense, the coefficients in $\hat{\mathbf{w}}$ have a clinical interpretation.

To summarize, each initial momentum can describe the local hippocampal shape changes *for a patient taken individually*, whereas the coefficient map $\hat{\mathbf{w}}$ can describe the relevance of hippocampal areas with regard to the disease progression, *at the population level* i.e. from the observation of *all training patients*.

4. Material and results

4.1. Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging, the National Instituteof Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public private partnership. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California — San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see http://www.adni-info.org.

A dataset of 206 hippocampus binary segmentations from 103 patients enrolled in ADNI (Mueller et al., 2005) has been used. The segmentations were computed and provided by ADNI, detailed information can be found on their website. For each patient, 'screening' and 'month 12' were the two time points selected. All patients were MCI at the screening point, 19 converted to AD by month 12, and the remaining 84 stayed MCI.

4.2. Experiments

4.2.1. Preprocessing

First, all screening images were resampled to a common isotropic voxel size $1.0\times1.0\times1.0$ mm, similar to their original size. Rigid transformations aligning the month 12 hippocampus towards the screening ones were computed using Ourselin et al. (2001).

4.2.2. Computation of initial momenta

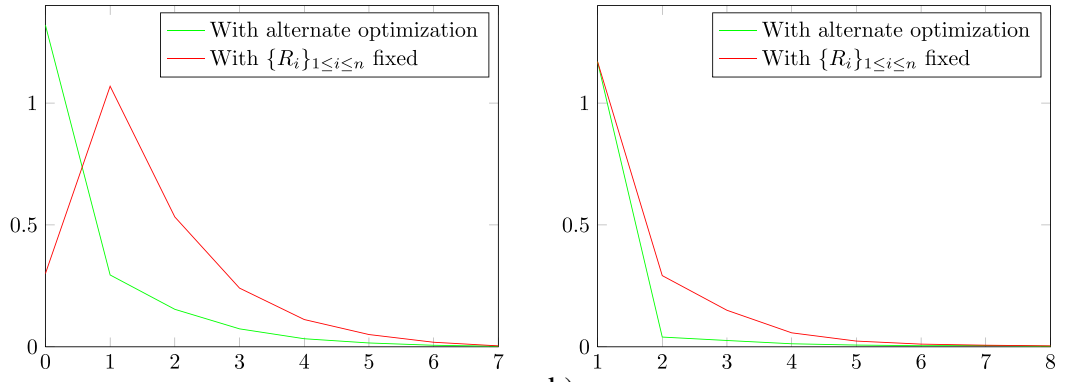
The geodesic shootings (Vialard et al., 2012a) were performed[3] using a sum of three kernels (sizes 1, 3 and 6 mm, with respective weights 2, 1 and 1), and 200 gradient descent iterations. To check the quality of the geodesic shooting computed for each patient i (second step in Fig. 2), the evolution of the Dice score DSC between S_t^i which is the deformed screening image at time t and the target image $F^i\circ(R^i)^{-1}$ was computed, and the average final DSC is 0.94 ± 0.01.

4.2.3. Computation of the template

The computation of a Karcher mean as described in Section 2.3 is a computationally expensive step, which is linear with the number of images. Therefore it can be desirable to select only a subset of the images. In this paper, a subset of 20 images was used, of corresponding hippocampal volumes which were the closest to the mean hippocampal volume. The Karcher mean estimate was updated four times, with respectively 200, 150, 150 and 100 gradient descent iterations in the geodesic shootings. Below are two verifications we performed to validate this approach.

First, we evaluated if all patients can be registered properly to the template, which is an important verification since only a subset of the images was used to compute the template. In our study, the average Dice score between the 103 registered patients and the template was 0.87 ± 0.02, which validated the suitability of the template obtained for our study. The last paragraph of Section 4.3 also provides another reason why such template can be used in our study.

[3] http://sourceforge.net/projects/utilzreg/.



a) Relative distance to the final estimate, i.e. $\frac{\|T_k - T_\infty\|^2_{L_2}}{\|T_\infty\|^2_{L_2}}$, where T_k is the Karcher estimate at iteration k, and T_∞ is approximated by the last computed estimate.

b) Relative distance between two consecutive estimates, i.e. $\frac{\|T_{k+1} - T_k\|^2_{L_2}}{\|T_k\|^2_{L_2}}$, where T_k is the Karcher estimate at iteration k.

Fig. 4. Empirical measures of convergence of the Karcher template algorithm. On this dataset, we notice that (1) the convergence speeds are coherent with the ones presented in Vialard et al. (2011) and Vialard et al. (2012b), i.e. only a few Karcher iterations are required for convergence, and (2) the alternate minimization over T and $\{R_i\}_{1 \leq i \leq n}$ provides a faster convergence than the one over T with the $\{R_i\}$ fixed.

Second, we evaluated the empirical convergence of our optimization procedure. Fig. 4a shows the relative distance to the final estimate, i.e.

$$\frac{\|T_k - T_\infty\|^2_{L_2}}{\|T_\infty\|^2_{L_2}}, \tag{23}$$

where T_k is the Karcher estimate at iteration k, and T_∞ is approximated by the last computed estimate. Fig. 4b shows the relative distance between two consecutive estimates, i.e.

$$\frac{\|T_{k+1} - T_k\|^2_{L_2}}{\|T_k\|^2_{L_2}}, \tag{24}$$

with the same notations. On this dataset, we notice that (1) the convergence speeds are coherent with the ones presented in Vialard et al. (2011, 2012b), i.e. only a few Karcher iterations are required for convergence, and (2) the alternate minimization over T and $\{R_i\}_{1 \leq i \leq n}$ provides a faster convergence than the one over T with the $\{R_i\}$ fixed.

4.2.4. Transport of initial momenta

To compute the transformations ϕ^i from the screening hippocampi towards the template (Fig. 3), rigid (Ourselin et al., 2001) then non-rigid (Modat et al., 2010) registration algorithms were applied with their default parameters. To check the quality of the registration ϕ^i computed to transport the local descriptor of the patient i (first step in 3), the Dice score was computed between the rigidly registered screening image and the template (i.e. $DSC(S \circ (R^i)^{-1}, T)$) and between the final registered screening image and the template (i.e. $DSC(S \circ (\phi^i)^{-1}, T)$).

4.2.5. Computation of the region of interest Ω_S

The region of interest Ω_S was restricted around the surface of the template (see Fig. 5), where the high values of the initial momenta lie. Moreover, this allows greater differences of coefficient values from one side to the other when using Sobolev regularization.

More specifically, given a binary template $T : \Omega \subset \mathbb{R}^3 \rightarrow [0,1]$ and a spherical structural element E_r of radius $r \in \mathbb{R}$ defined as

$$E_r \overset{\text{def.}}{=} \left\{(\omega_1, \omega_2, \omega_3) \in \mathbb{R}^3;\ \omega_1^2 + \omega_2^2 + \omega_3^2 \leq r^2\right\}, \tag{25}$$

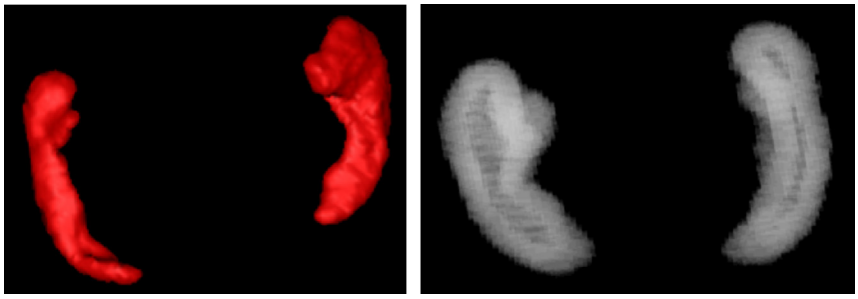
we define the region of interest Ω_S as

$$\Omega_S \overset{\text{def.}}{=} \text{Dila}(T, E_r) - \text{Ero}(T, E_r), \tag{26}$$

where Dila and Ero are the standard dilatation and erosion morphological operators. In this study, using $r = 5$, the ROI Ω_S contained 12,531 voxels.

4.2.6. Optimization of the logistic regression model

In the training procedure, we have $n = 103$ observations (one for each patient). As initial momenta are scalar fields in space, each initial momenta has the same dimension as the number of voxels, so $p = 12{,}531$. Since stable and progressive classes in the dataset are



a) Template T b) Region of interest Ω_S

Fig. 5. The region of interest Ω_S (visualized with transparency) is designed to select voxels close to the boundary (i.e. close to the surface) of T. It is obtained via standard morphological operations, and in this study Ω_S contains 12,531 voxels.

unbalanced, the weighted version of the loss function defined in Section 3.2.3 was used. Solution of the optimization problems was computed via HANSO[4] with a maximum of 20 iterations.

4.2.7. Performance evaluation

First, the effect of spatial regularizations was compared. The spatial regularizations introduced in Section 3.2 aim at enforcing local correlations between the coefficients in $\mathbf{w}$. Using the whole dataset, the effects of the various regularizations were compared. Second, the model was evaluated in terms of classification of AD progression. All patients were classified using a leave-10%-out scheme. From the numbers of true/false positives/negatives (TP, FP, TN, FN), four indicators were used to measure classification accuracy: specificity $Spec \overset{\text{def.}}{=} \frac{TN}{TN+FP}$, sensitivity $Sens \overset{\text{def.}}{=} \frac{TP}{TP+FN}$, negative predictive value $NPV \overset{\text{def.}}{=} \frac{TN}{TN+FN}$, and positive predictive value $PPV \overset{\text{def.}}{=} \frac{TP}{TP+FP}$. Statistical tests were also performed to evaluate the significance of the differences. Using $N = 50$ random re-orderings of the patients, the Spec + Sens variable was computed 50 times for each regularization and two-sample t-tests were performed.

4.3. Effect of spatial regularizations

When using standard regularizations, increasing the regularization does not lead to any spatial coherence (Fig. 6a, b and c). It is interesting to remark that LASSO regularization emphasizes a more limited number of points than ridge regularization. This is particularly clear in the right columns of Fig. 6, where the regularization energy ($\lambda J(\mathbf{w})$ in Eq. (16)) has a significant weight in the total energy. As expected, ElasticNet also gives results which are in-between those of LASSO and those of ridge. In contrast to these regularization techniques, the higher the spatial regularizations, the more structured are the coefficients. Note that delimited areas are coherent across different spatial regularizations. Sobolev regularization leads to smooth coefficient maps (Fig. 6d) whereas total variation tends to piecewise constant maps (Fig. 6e). Finally, fused LASSO adds sparsity by zeroing out the lowest coefficients (Fig. 6f).

4.3.1. Another benefit of spatial regularizations

As mentioned in the Introduction, a motivation to regularize the learning problem is the low number of observations compared to the dimensionality of the problem. However, we can infer another benefit of the use of spatial regularizations. Indeed, to build voxel-based statistical models from the observations of several patients, one needs to align these observations properly. Even though we checked the quality of the alignment to the template, such alignment is not perfect. Adding spatial regularizations in the model is a way to limit the effects of the alignment errors.

4.4. Classification of Alzheimer's disease progression

Besides providing a map of coefficients indicating the importance of each voxel with regard to the disease progression, the model presented in this paper can be used to classify the disease progression of new patients. Table 1 displays the classification performance indicators of binary classification using logistic loss and various regularizations.

Without any regularization, the resulting classifier always predicts the same class. Before going any further, let us comment on this point. If all testing subjects are classified in the same class, it means that all the testing points are on the same side on the hyperplane found in the optimization process. Here, unbalanced observations and the chosen optimization strategy are the causes of this result. In the model used, the bias b plays a special role and several strategies can be considered, such as 1) optimizing $\mathbf{w}$ and b at the same time, 2) optimizing $\mathbf{w}$ and b, then freezing $\mathbf{w}$ and optimizing b, 3) optimizing $\mathbf{w}$ and b, then freezing $\mathbf{w}$ and setting b using heuristic rules (e.g. setting it to have the same ratio between classes in training and test sets), 4) optimizing $\mathbf{w}$ with b frozen to zero, then optimizing b, 5) optimizing $\mathbf{w}$ with b frozen to zero, then setting b using heuristic rules, etc. In initial tests, we realized that some strategies would classify all patients to positive whereas other would classify them all to negative. This happened when the optimization is not regularized. However, this instability with regard to the optimization strategy fades out when the problem is regularized. These initial tests further motivated the use of regularization. Let us note that the above strategy 1) was used in all the results presented in this paper.

All regularizations improve significantly the classification performance, the top 3 being the three spatial regularizations. On this dataset, fused LASSO is the one providing the best results (Spec + Sens = 1.32), closely followed by total variation (Spec + Sens = 1.31).

4.4.1. Comparison with the literature

Using spatial regularizations such as total variation and fused-LASSO, our experiments provide higher performances than the best one reported in Fiot et al. (2012) (Spec + Sens = 1.27). Moreover, the linear classification model used in this paper is simpler than the non-linear SVM used in Fiot et al. (2012). SVM is a very powerful approach, which has been widely studied and successfully used. Many implementations are available, but it can get difficult to modify them and, for example, add spatial regularizations. Besides, only linear SVM can provide an interpretable map of coefficients, but not the non-linear version used in Fiot et al. (2012). On the other hand, a model as simple as the logistic regression can be easily implemented and modified.

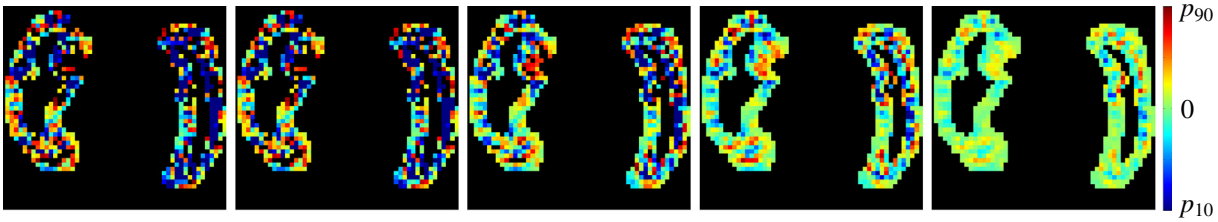
4.5. Statistical tests

To evaluate the significance of the performance differences found in Table 1, we performed two-sample t-tests. The variable considered was Spec + Sens, and 50 realizations of the variable from random re-ordering of the patients were obtained for each sample. Two regularizations can be considered statistically significantly different if the test has a p-value $p < \alpha = 10^{-3}$. These results are presented in Table 2. First, we notice that all regularizations are statistically better than the absence of regularization. Then we notice that all spatial regularizations are statistically better than standard regularizations. Finally, we notice that despite higher prediction accuracy, Elastic Net is not statistically significantly better than ridge in our tests. Similarly, fused-LASSO is not statistically significantly better than total variation in our tests.

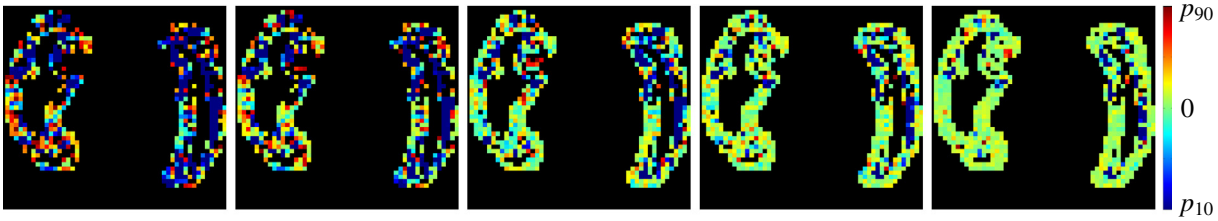
4.6. Computation time

The various algorithms were implemented in a mix of C++, MATLAB®, mex and python. Table 3 reports approximate running time on a standard laptop (Intel® Core™ i7-2720QM CPU at 2.20 GHz, 8 GB of RAM). The geodesic shooting step is linear with the number of patients. The computation of the template is linear with the number of patients and the number of Karcher iterations. One should note that Karcher iterations can have decreasing number of gradient descent iterations, which decreases the total computation time. Then the transport is also linear with the number of patients. So far, it is interesting to notice that all the steps can be easily be divided into different jobs to take advantage of multi-core or distributed architectures. Finally come the learning and classification. The computation time of this step can vary dramatically depending on several parameters such as the training/testing splitting scheme, the optimization algorithm, and the number of regularization parameters to test. In particular, for this exploratory study, we used mainly HANSO algorithm, since the convergence rate of the proximal algorithms mentioned in Section 3.2.2 was too low.

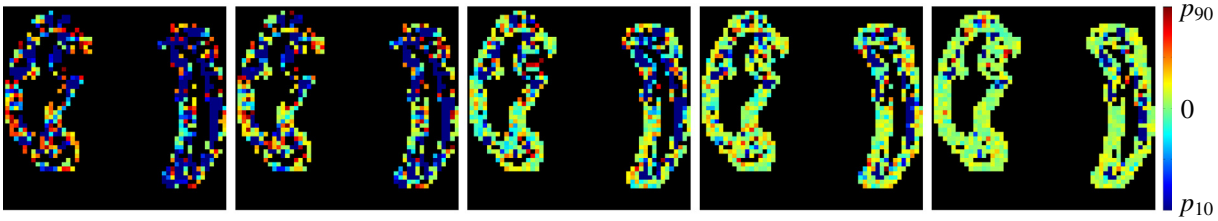
[4] http://www.cs.nyu.edu/overton/software/hanso.



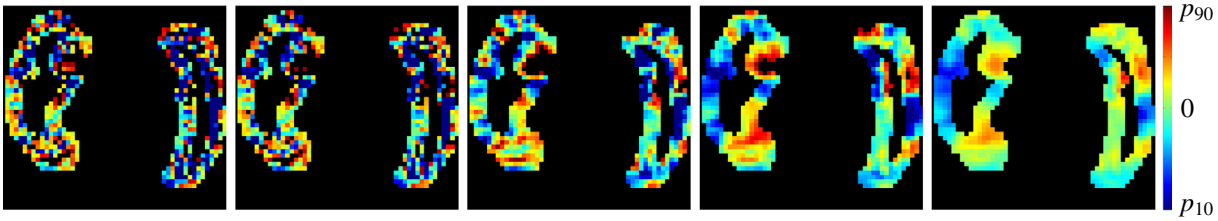
(a) Ridge



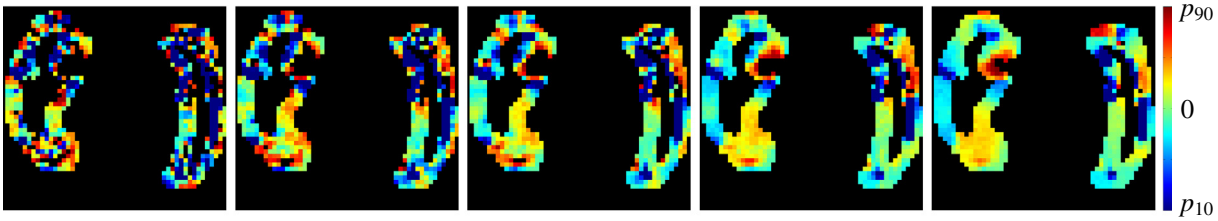
(b) LASSO



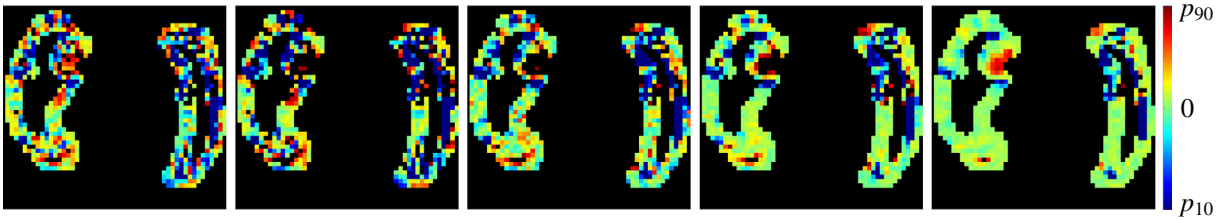
(c) Elastic Net



(d) Sobolev



(e) Total variation



(f) Fused LASSO

Fig. 6. Effects of various regularizations on the solution $\hat{\mathbf{w}}$ of the optimization problem. Each small image represents the coefficients of one 2D slice of $\hat{\mathbf{w}}$, which is a 3D volume. Zero coefficients are displayed in light green, higher values are going red and lower values are going blue. On each row, the regularization is increasing from left to right, and the 10th and 90th percentiles of the coefficients (resp. P_{10} and P_{90}) correspond to the saturation limits of the colorbar. Panels a, b and c show standard regularizations whereas Panels d, e and f show spatial regularizations. Spatial regularizations provide more structured coefficients.

Table 1
Prediction accuracy of MCI patients' progression.

Regularization		λ range	$\hat{\lambda}$ (optimal λ)	Spec+ Sens	Spec	Sens	NPV	PPV
None		0	0	1.00	0.00	1.00	NaN	0.18
Standard	LASSO	$[10^{-9},10^{0}]$	0.01	1.04	0.20	0.84	0.85	0.19
	Ridge	$[10^{-9},10^{0}]$	0.001	1.06	0.95	0.11	0.82	0.33
	Elastic Net	$[10^{-9},10^{0}]^2$	$\begin{cases}\hat{\lambda}_1 = 0.01\\ \hat{\lambda}_2 = 1\end{cases}$	1.13	0.29	0.84	0.89	0.2
Spatial	Sobolev	$[10^{-9},10^{7}]$	10^{4}	1.17	0.54	0.63	0.87	0.24
	Total Variation	$[10^{-9},10^{0}]$	0.01	**1.31**	**0.46**	**0.84**	**0.93**	**0.26**
	Fused LASSO	$[10^{-9},10^{0}]^2$	$\begin{cases}\hat{\lambda}_1 = 0.01\\ \hat{\lambda}_2 = 10^{-4}\end{cases}$	**1.32**	**0.48**	**0.84**	**0.93**	**0.27**

4.7. Comparison with the literature

As mentioned earlier, the main contribution of this paper is the comparison of the effects of various regularizations on the solution of binary classification problem with a logistic loss. In the context of longitudinal Alzheimer's disease study, we saw that the use of spatial regularizations techniques was not only leading to better classification results than standard regularizations, but also providing maps of coefficients with improved spatial coherence.

In the literature, a large number of methods are also trying to identify the hippocampal sub-areas that are related to either the conversion of patients to the disease or to other symptoms such as cognitive or memory measures. For example, one can cite Fig. 5 of Frisoni et al. (2008), Fig. 7 of Gutman et al. (2009), Fig. 1 to 5 of Apostolova et al. (2010), and Fig. 3 and 4 of Shen et al. (2012).

Several strategies can be considered to compare the most significant regions found by various methods. One strategy is to transport relevance maps from different methods to the same space. However, transporting information is delicate (Fiot et al., 2012), and one needs to be cautious with such strategy. This transport could be avoided by using the same template for all methods, though this is likely to cause problems if the population studied is not the same. Another strategy is to rank the hippocampal subareas, as it is done for example in Table 2 of Frisoni et al. (2008), and compare the rankings. This strategy would require us to align a map of known hippocampal subareas to our template, and design a ranking algorithm (for example based on $\int_{\omega\in\Omega_R}\hat{\mathbf{w}}(\omega)^2 d\omega$, where Ω_R is a hippocampal subregion).

Comparing qualitatively or quantitatively the subregions that are the most significant with regard to disease progression is out of the scope of this paper. Nonetheless, it is a very interesting perspective, and several strategies including the ones mentioned above are considered for future work.

5. Conclusion

In this paper, we studied deformation models for longitudinal population analysis, regularizations and machine learning strategies. In particular, we investigated the combined use of the LDDMM framework and classification with logistic loss and spatial regularizations in the context of Alzheimer's disease. Results indicate that initial momenta of hippocampus deformations are able to capture information relevant to the progression of the disease.

Another contribution of this paper is the joint use of a simple linear classifier with complex spatial regularizations. Achieving results higher than the ones reported in Fiot et al. (2012), which uses non-linear SVM classifier, our method provides in addition coefficient maps with direct anatomical interpretation.

Moreover, we compared Sobolev, total variation and fused LASSO regularizations. While they all successfully enforce different priors (respectively smooth, piecewise constant and sparse), their resulting coefficient maps are coherent one to the other. They improve coefficient maps and their classification performances are statistically better than the ones obtained with standard regularizations.

Now the ideas and results presented in this paper open a wide range of perspectives. First, the question of the representation of patients from images, and in particular the representation of their evolutions for longitudinal population studies was raised. We have used initial momenta encoding the patient evolution in 3D volumes. An interesting research direction is the adaptation of our pipeline to surface representation of shape evolution. Indeed, as we saw in the application studied in this paper, the strong values of the initial momenta lie on the hippocampus volume boundary, in other words on the surface. Second, the question of how to compare evolutions of different patients was raised. We studied the use of Karcher mean and the importance of the regularizations. Even though diffeomorphic deformation models such as LDDMM can provide smooth deformation fields and encode the shape deformation of a

Table 2
Statistical p-values of two-sample t-tests between different regularizations. The variable considered is Spec + Sens, and 50 realizations of the variable from random re-orderings of the patients were obtained for each sample. Two regularizations can be considered statistically significantly different if the test has a p-value light green $p < \alpha = 10^{-3}$ (marked in green, red otherwise).

Regularization		Standard			Spatial		
		LASSO	Ridge	Elastic Net	Sobolev	Total Variation	Fused LASSO
None		$<10^{-5}$	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$
Standard	LASSO	–	$1.1*10^{-04}$	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$
	Ridge		–	$4.2*10^{-02}$	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$
	Elastic Net			–	$6.3*10^{-05}$	$<10^{-5}$	$<10^{-5}$
Spatial	Sobolev				–	$<10^{-5}$	$<10^{-5}$
	Total Variation					–	0.86
	Fused LASSO						–

Table 3
Computation time of the various steps. (*): can differ by several orders of magnitude, see Section 4.6 for details.

Step	Computation time
Preprocessing	A few hours
Geodesic shooting	≈1 day
Template computation	≈3 days
Transport	≈1 day
Learning and classification	From 1 min to several days*

patient in a smooth representation, we saw that it is important to regularize spatially across the population (i.e. between patients) in order to be able to build meaningful statistical models for classification and biomarker discovery. On that point, the logistic regression model has proven to be efficient as it can be combined with complex regularizations. Our spatial regularizations gave the best results on our dataset, and another research direction is the study of other regularizations such as group sparsity. Third, another great perspective of this work consists in studying evolutions of patients with more than two time points. In this context, the design of spatio-temporal regularizations (for example in the context of geodesic regression (Niethammer et al., 2011)) is an exciting research direction.

Acknowledgements

Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec, Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Allassonnière, S., Kuhn, E., Trouvé, A., 2008. 2008. MAP estimation of statistical deformable templates via nonlinear mixed effects models: deterministic and stochastic approaches. (Session 03: Building Atlases) In: Pennec, X. (Ed.), 2nd MICCAI Workshop on Mathematical Foundations of Computational Anatomy, New-York, United States, pp. 80–91.

Apostolova, L.G., Morra, J.H., Green, A.E., Hwang, K.S., Avedissian, C., Woo, E., Cummings, J.L., Toga, A.W., Jack, C.R., Weiner, M.W., Thompson, P.M., Initiative, A.D.N., 2010. Automated 3D mapping of baseline and 12-month associations between three verbal memory measures and hippocampal atrophy in 490 ADNI subjects. NeuroImage 51, 488–499.

Ashburner, J., Friston, K.J., 2011. Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation. NeuroImage 55, 954–967.

Avants, B., Gee, J.C., 2004. Geodesic estimation for large deformation anatomical shape averaging and interpolation. NeuroImage 23, S139–S150 (Mathematics in Brain Imaging).

Baldassarre, L., Mourao-Miranda, J., Pontil, M., 2012. Structured sparsity models for brain decoding from FMRI data. Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop, pp. 5–8. http://dx.doi.org/10.1109/PRNI.2012.31.

Beck, A., Teboulle, M., 2009. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Trans. Image Process. 18, 2419–2434.

Beg, M.F., Khan, A., 2006. Computing an average anatomical atlas using LDDMM and geodesic shooting. Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on, IEEE, pp. 1116–1119.

Beg, M.F., Miller, M.I., Trouvé, A., Younes, L., 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. 61, 139–157.

Bhatia, K.K., Hajnal, J.V., Puri, B.K., Edwards, A.D., Rueckert, D., 2004. Consistent groupwise non-rigid registration for atlas construction. Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on, vol. 1, pp. 908–911. http://dx.doi.org/10.1109/ISBI.2004.1398686.

Bossa, M.N., Zacur, E., Olmos Gasso, S., 2010. On changing coordinate systems for longitudinal tensor-based morphometry. Medical Image Computing and Computer-assisted Intervention (MICCAI): Intl. Workshop of Spatio-temporal Image Analysis for Longitudinal and Time-series Image Data (STIA).

Chupin, M., Mukuna-Bantumbakulu, A.R., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnéhun, S., Lemieux, L., Dubois, B., Garnero, L., 2007. Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: method and validation on controls and patients with Alzheimer's disease. NeuroImage 34, 996–1019. http://dx.doi.org/10.1016/j.neuroimage.2006.10.035.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., Initiative, A.D.N., 2009. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19, 579–587. http://dx.doi.org/10.1002/hipo.20626.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56, 766–781. http://dx.doi.org/10.1016/j.neuroimage.2010.06.013.

Cuingnet, R., Glaunès, J.A., Chupin, M., Benali, H., Colliot, O., 2012. Spatial and anatomical regularization of SVM: a general framework for neuroimaging data. IEEE Trans. Pattern Anal. Mach. Intell. http://dx.doi.org/10.1109/TPAMI.2012.142.

Desikan, R.S., Cabral, H.J., Hess, C.P., Dillon, W.P., Glastonbury, C.M., Weiner, M.W., Schmansky, N.J., Greve, D.N., Salat, D.H., Buckner, R.L., Fischl, B., Initiative, A.D.N., 2009. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain 132, 2048–2057.

Durrleman, S., Allassonnière, S., Joshi, S., 2013. Sparse adaptive parameterization of variability in image ensembles. Int. J. Comput. Vis. 101, 161–183. http://dx.doi.org/10.1007/s11263-012-0556-1.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imaging 26, 93–105. http://dx.doi.org/10.1109/TMI.2006.886812.

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., Initiative, A.D.N., 2008a. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. NeuroImage 39, 1731–1743.

Fan, Y., Resnick, S.M., Wu, X., Davatzikos, C., 2008b. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. NeuroImage 41, 277–285. http://dx.doi.org/10.1016/j.neuroimage.2008.02.043.

Fiot, J.B., Risser, L., Cohen, L.D., Fripp, J., Vialard, F.X., 2012. Local vs global descriptors of hippocampus shape evolution for Alzheimer's longitudinal population analysis. 2nd International MICCAI Workshop on Spatiotemporal Image Analysis for Longitudinal and Time-series Image Data (STIA '12), Nice, France, pp. 13–24. http://dx.doi.org/10.1007/978-3-642-33555-6.

Fletcher, P.T., Lu, C., Pizer, M., Joshi, S., 2004. Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Med. Imaging 995–1005.

Fréchet, M., 1948. Les éléments aléatoires de nature quelconque dans un espace distancie. Ann. Inst. Henri Poincaré 10, 215–310.

Frisoni, G.B., Ganzola, R., Canu, E., Rüb, U., Pizzini, F.B., Alessandrini, F., Zoccatelli, G., Beltramello, A., Caltagirone, C., Thompson, P.M., 2008. Mapping local hippocampal changes in Alzheimer's disease and normal ageing with mri at 3 tesla. Brain 131, 3266–3276. http://dx.doi.org/10.1093/brain/awn280.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., Eustache, F., Colliot, O., Initiative, A.D.N., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. NeuroImage 47, 1476–1486.

Gramfort, A., Thirion, B., Varoquaux, G., 2013. Identifying predictive regions from fMRI with TV-L1 prior. Pattern Recognition in Neuroimaging (PRNI), IEEE, Philadelphia, United States. ANR grant BrainPedia, ANR-10-JCJC 1408-01, FMJH Program Gaspard Monge in Optimization and Operation Research with Support from EDF.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with GraphNet. NeuroImage 72, 304–321.

Gutman, B., Wang, Y., Morra, J., Toga, A.W., Thompson, P.M., 2009. Disease classification with hippocampal shape invariants. Hippocampus 19, 572–578. http://dx.doi.org/10.1002/hipo.20627.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67. http://dx.doi.org/10.2307/1267351.

Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Eger, E., Bach, F., Thirion, B., 2012. Multiscale mining of fMRI data with hierarchical structured sparsity. SIAM J. Imaging Sci. 5, 835–856.

Jia, H., Wu, G., Wang, Q., Shen, D., 2010. Absorb: atlas building by self-organized registration and bundling. NeuroImage 51, 1057–1070. http://dx.doi.org/10.1016/j.neuroimage.2010.03.010.

Joshi, S., Davis, B., Jomier, M., Gerig, G., 2004. Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131, 681–689. http://dx.doi.org/10.1093/brain/awm319.

Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S.M., Davatzikos, C., 2004. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. NeuroImage 21, 46–57.

Lewis, A.S., Overton, M.L., 2012. Nonsmooth optimization via quasi-Newton methods. Math. Program. 1–29. http://dx.doi.org/10.1007/s10107-012-0514-2.

Ma, J., Miller, M.I., Trouvé, A., Younes, L., 2008. Bayesian template estimation in computational anatomy. NeuroImage 42, 252–261. http://dx.doi.org/10.1016/j.neuroimage.2008.03.056.

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehéricy, S., Benali, H., 2009. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51, 73–83. http://dx.doi.org/10.1007/s00234-008-0463-x.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B., 2011. Total variation regularization for fMRI-based prediction of behaviour. IEEE Trans. Med. Imaging 30, 1328–1340. http://dx.doi.org/10.1109/TMI.2011.2113378.

Miller, M.I., Trouvé, A., Younes, L., 2006. Geodesic shooting for computational anatomy. J. Math. Imaging Vis. 24, 209–228. http://dx.doi.org/10.1007/s10851-005-3624-0.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. 98, 278–284. http://dx.doi.org/10.1016/j.cmpb.2009.09.002.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. 15, 869–877. http://dx.doi.org/10.1016/j.nic.2005.09.008 (Alzheimer's Disease: 100 Years of Progress).

Ng, B., Abugharbieh, R., 2011. Generalized sparse regularization with application to fMRI brain decoding. In: Székely, G., Hahn, H. (Eds.), Information processing in medical imaging. Volume 6801 of Lecture Notes in Computer Science, vol. 6801. Springer, Berlin Heidelberg, pp. 612–623. http://dx.doi.org/10.1007/978-3-642-22092-0_50.

Niethammer, M., Huang, Y., Vialard, F.X., 2011. Geodesic regression for image time-series. Med. Image Comput. Comput. Assist Interv. 14, 655–662.

Ourselin, S., Roche, A., Subsol, G., Pennec, X., Ayache, N., 2001. Reconstructing a 3D structure from serial histological sections. Image Vis. Comput. 19, 25–31. http://dx.doi.org/10.1016/S0262-8856(00)00052-4.

Pennec, X., 1999. Probabilities and statistics on Riemannian manifolds: basic tools for geometric measurements. In: Cetin, A., Akarun, L., Ertuzun, A., Gurcan, M., Yardimci, Y. (Eds.), Proc. of Nonlinear Signal and Image Processing (NSIP '99), IEEE-EURASIP, June 20–23, Antalya, Turkey, pp. 194–198.

Pennec, X., 2006. Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. J. Math. Imaging Vis. 25, 127–154. http://dx.doi.org/10.1007/s10851-006-6228-4.

Qiu, A., Younes, L., Miller, M.I., Csernansky, J.G., 2008. Parallel transport in diffeomorphisms distinguishes the time-dependent pattern of hippocampal surface deformation due to healthy aging and the dementia of the Alzheimer's type. NeuroImage 40, 68–76.

Querbes, O., Aubry, F., Pariente, J., Lotterie, J.A., Démonet, J.F., Duret, V., Puel, M., Berry, I., Fort, J.C., Celsis, P., Initiative, A.D.N., 2009. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 132, 2036–2047.

Raguet, H., Fadili, J., Peyré, G., 2013. A generalized forward–backward splitting. SIAM J. Imaging Sci. 6, 1199–1226. http://dx.doi.org/10.1137/120872802.

Risser, L., Vialard, F.X., Wolz, R., Murgasova, M., Holm, D.D., Rueckert, D., 2011. Simultaneous multiscale registration using large deformation diffeomorphic metric mapping. IEEE Trans. Med. Imaging 30 (10), 1746–1759.

Seghers, D., D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P., 2004. Construction of a brain template from MR images using state-of-the-art registration and segmentation techniques. In: Barillot, C., Haynor, D., Hellier, P. (Eds.), Medical image computing and computer-assisted intervention — MICCAI 2004. Volume 3216 of Lecture Notes in Computer Science, vol. 3216. Springer, Berlin Heidelberg, pp. 696–703. http://dx.doi.org/10.1007/978-3-540-30135-6_85.

Shen, K.K., Fripp, J., Mériaudeau, F., Chételat, G., Salvado, O., Bourgeat, P., A.D.N Initiative, 2012. Detecting global and local hippocampal shape changes in Alzheimer's disease using statistical shape models. NeuroImage 59, 2155–2166. http://dx.doi.org/10.1016/j.neuroimage.2011.10.014.

Singh, N., Fletcher, P., Preston, J., Ha, L., King, R., Marron, J., Wiener, M., Joshi, S., 2010. Multivariate statistical analysis of deformation momenta relating anatomical shape to neuropsychological measures. In: Jiang, T., Navab, N., Pluim, J., Viergever, M. (Eds.), Medical image computing and computer-assisted intervention — MICCAI 2010. Volume 6363 of Lecture Notes in Computer Science, vol. 6363. Springer, Berlin/Heidelberg, pp. 529–537.

Tibshirani, R., 1994. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288.

Vaillant, M., Miller, M.I., Younes, L., Trouvé, A., 2004. Statistics on diffeomorphisms via tangent space representations. NeuroImage 23 (Suppl. 1), S161–S169. http://dx.doi.org/10.1016/j.neuroimage.2004.07.023.

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen Jr., R.C., C.R.J., 2008. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. NeuroImage 39, 1186–1197. http://dx.doi.org/10.1016/j.neuroimage.2007.09.073.

Vialard, F.X., Risser, L., Holm, D., Rueckert, D., 2011. Diffeomorphic Atlas Estimation using Karcher Mean and Geodesic Shooting on Volumetric Images.

Vialard, F.X., Risser, L., Rueckert, D., Cotter, C.J., 2012a. Diffeomorphic 3D Image Registration via Geodesic Shooting using an Efficient Adjoint Calculation. Int. J. Comput. Vis. 97, 229–241. http://dx.doi.org/10.1007/s11263-011-0481-8.

Vialard, F.X., Risser, L., Rueckert, D., Holm, D., 2012b. Diffeomorphic Atlas Estimation using Karcher Mean and Geodesic Shooting on Volumetric Images. Annals of the British Machine Vision Association.

Wang, L., Beg, F., Ratnanather, T., Ceritoglu, C., Younes, L., Morris, J.C., Csernansky, J.G., Miller, M.I., 2007. Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type. IEEE Trans. Med. Imaging 26, 462–470. http://dx.doi.org/10.1109/TMI.2005.853923.

Younes, L., 2007. Jacobi fields in groups of diffeomorphisms and applications. Q. Appl. Math. 65, 113–134.

Yushkevich, P.A., Avants, B.B., Das, S.R., Pluta, J., Altinay, M., Craige, C., 2010. Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: an illustration in ADNI 3 T MRI data. NeuroImage 50, 434–445.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.